ISSN 1755-5361

University of Essex

Department of Economics

Discussion Paper Series

No. 695 January 2011

Poisson: Some convergence issues

J.M.C. Santos Silva & Silvana Tenreyro

`poisson`: Some convergence issues*

J.M.C. Santos Silva† Silvana Tenreyro‡

January 6, 2011

Abstract

In this note we identify and illustrate some shortcomings of the `poisson` command in STATA. Specifically, we point out that the command fails to check for the existence of the estimates and show that it is very sensitive to numerical problems. While these are serious problems that may prevent users from obtaining estimates, or even produce spurious and misleading results, we show that the informed user often has available simple work-arounds for these problems.

Key words: Collinearity, Complete separation, Numerical problems, Perfect prediction.

*We are very grateful to an anonymous referee for most valuable comments and suggestions. We also thank Styliani Christodoulopoulou for helpful discussions and comments on an earlier version of the paper. The usual disclaimer applies. Santos Silva gratefully acknowledges partial financial support from Fundação para a Ciência e Tecnologia (FEDER/POCI 2010).

†University of Essex and CEMAPRE. Wivenhoe Park, Colchester CO4 3SQ, United Kingdom. Fax: +44 (0)1206 872724. E-mail: jmcss@essex.ac.uk.

‡London School of Economics, CREI, CEP, and CEPR. Department of Economics, s.600. St. Clement's Building. Houghton St., London WC2A 2AE, United Kingdom. Fax: +44 (0)20 78311840. E-mail: s.tenreyro@lse.ac.uk.

1. INTRODUCTION

Besides being the most widely used estimator for count data (see Winkelmann, 2008, and Cameron and Trivedi, 1997), Poisson regression is also becoming increasingly used to estimate multiplicative models for other non-negative data (see, among others, Manning and Mullahy, 2001, and Santos Silva and Tenreyro, 2006). The availability in STATA of a command that estimates Poisson regression has been an important reason for the increasing popularity of this estimator. However, researchers using Poisson regression, especially those using it to estimate gravity equations as recommended by Santos Silva and Tenreyro (2006), often find that the algorithm implemented in STATA's `poisson` command does not converge. There are two main reasons for this. First, as noted by Santos Silva and Tenreyro (2010), there are instances in which the estimates do not exist and if that is the case the convergence of the algorithm used to maximize the likelihood function can only be spurious. Second, even when the estimates exist, researchers using STATA may have trouble to get Poisson regression estimates because the `poisson` command is very sensitive to numerical problems. In this note we describe how researchers can identify some of the situations that may lead to convergence problems and propose some simple work-arounds.

2. THE NON EXISTENCE OF THE ESTIMATES

Let y_i and x_i denote respectively the variate of interest and the vector of covariates, and assume that the researcher specifies $\mathrm{E}(y_i|x_i) = \exp(x_i'\beta)$. In a sample of size n, $\hat{\beta}$, the Poisson regression estimate of β, is defined by

$$\sum_{i=1}^{n} \left[y_i - \exp\left(x_i'\hat{\beta}\right) \right] x_i = 0. \tag{1}$$

The form of (1) makes clear that β will be consistently estimated as long as the conditional mean is correctly specified. That is, the only condition required for the consistency of the estimator is that $\mathrm{E}(y_i|x_i) = \exp(x_i'\beta)$. This is the well known pseudo-maximum likelihood result of Gourieroux, Monfort and Trognon (1984).

However, Santos Silva and Tenreyro (2010) have shown that $\hat{\beta}$ does not always exist and that its existence depends on the data configuration. In particular, the estimates may not exist if there is perfect collinearity for the sub-sample with positive observations of y_i.[1] If the estimates do not exist, it is either impossible for the estimation algorithm to converge or convergence is spurious. The following STATA code illustrates the situation where convergence is not achieved:[2]

```
drawnorm x1, n(1000) seed(101010) double clear
generate double u=rpoisson(1)
generate y=exp(1+10*x1)*u
generate double x2=(y==0)
poisson y x1 x2, robust
```

An example where the convergence is spurious is given by the code below:

```
drawnorm x1, n(1000) seed(101010) double clear
generate double y=rpoisson(1)
generate double x2=(y==0)
poisson y x1 x2, robust
```

The non-existence of the maximum likelihood estimates in Poisson regression is analogous to what happens in binary choice models when there is complete separation or quasi-complete separation, as described by Albert and Anderson (1984) and Santner and Duffy (1986). In the case of binary models, it is standard to check for the existence of the estimates before starting the actual estimation. In contradistinction, the `poisson` command in STATA does not check for the existence of the estimates and therefore it is important that users investigate whether or not the estimates exist. Because the regressors that may cause the non-existence of the estimates are characterized by their perfect collinearity with the others for the sub-sample with $y_i > 0$, they can easily be identified in STATA by using a simplified version of the three-step method suggested by Santos Silva and Tenreyro (2010):

[1] See also Haberman (1973).

[2] The code used in this note produces the desired results in STATA/IC 11.1 for Windows (32-bit). Using other flavours of STATA, for example MP versions, may lead to different outcomes.

Step 1: Construct a subset of explanatory variable, say $\tilde{x}_i$, comprising only the regressors that are not collinear for the observations with $y_i > 0$;

Step 2: Using the full sample, run the Poisson regression of y_i on $\tilde{x}_i$.

The following code, where it is assumed that all variables with names starting with `x` are regressors, illustrates the implementation of the procedure:[3]

```
local _rhs "x*"
_rmcoll `_rhs' if y>0
poisson y `r(varlist)', robust
```

This procedure ensures that the estimates exist by eliminating all potentially problematic regressors, even those that actually do not lead to the non-existence of the maximum likelihood estimates.[4] Therefore, the researcher should subsequently investigate one-by-one all the variables that were dropped to see if any of them can be included in the model. Carefully investigation of the variables to be excluded is particularly important when the model contains sets of dummies with several categories: in this case dropping one of the dummies implies an arbitrary redefinition of the reference category which is unlikely to be sensible. In any case, dropping some regressors should never be an automatic procedure because it changes the model specification and therefore the researcher should carefully consider what is the best way to find an interesting specification for which the (pseudo) maximum likelihood estimates exist.

It is worth noting that the non existence of the estimates can also occur in any regression model where the conditional mean is specified in such a way that its image does not include all the points in the support of the dependent variable. Therefore, unless the data are strictly positive, this problem can occur not only in the Poisson regression but also in other models specifying $\mathrm{E}(y_i|x_i) = \exp(x_i'\beta)$, and in models for limited dependent variables like the Tobit (Tobin, 1958). In all these cases, the identification of the problematic regressors can be done using methods akin to the one described above.

[3] We are grateful to Markus Baldauf for help with the development of an earlier version of this code and to an anonymous referee for suggesting this much simple version using the `_rmcoll` command.

[4] A less strict criterion to select the regressors to be dropped is used by default in the `ppml` command briefly discussed below.

3. NUMERICAL DIFFICULTIES

Even if the (pseudo) maximum likelihood estimates of the Poisson regression exist, STATA may have difficulty identifying them due to the sensitivity to numerical problems of the algorithms available in the `poisson` command. In particular, we are aware of three situations in which the algorithms in the `poisson` command have trouble locating the maximum and may not converge, even when the (pseudo) maximum likelihood estimates of the Poisson regression are well defined.

The simplest case in which STATA finds it difficult to find the Poisson (pseudo) maximum likelihood estimates is when y has some very large values. The following STATA code illustrates the situation:[5]

```
drawnorm u x1 x2, n(1000) seed(101010) double clear
generate double y=exp(40+x1+x2+u)
poisson y x1 x2, robust difficult
```

In this example the Poisson regression does not converge, at least not in a reasonable number of iterations. Obviously, in this case the problem can easily be by-passed just by re-scaling the dependent variable, say, by dividing it by $\exp(40)$.

A second situation in which STATA finds it difficult to locate the solution of (1) occurs when the regressors are highly collinear and have very different magnitudes. The following STATA code illustrates the situation:[6]

```
drawnorm u e x1, n(1000) seed(101010) double clear
generate x2=(x1<-2)
generate double x3=20+x1+(e/100)*(x1<-2)
generate double y=exp(1+x1+x2+u)
poisson y x1 x2 x3, robust difficult
```

In this case, again, the Poisson regression does not converge but a simple work-around is available: if the third regressor is re-centered at zero convergence is achieved with ease.

These two examples suggest that, when facing convergence problems, researchers should re-scale and re-center their data in a way that reduces possible numerical problems. However, even if that is done, STATA will have trouble finding the (pseudo) maximum like-

[5] We are grateful to Alexandros Theloudis for showing us a dataset where this situation occurs.

[6] We are grateful to Avni Hanedar for showing us a dataset where this situation occurs.

lihood estimates of the Poisson regression when the covariates are extremely (but not perfectly) collinear. The following example illustrates this situation:

```
drawnorm u e x1, n(1000) seed(101010) double clear
generate double x2=(x1+e/20000)
generate double y=exp(1+x1+x2+u)
poisson y x1 x2, robust difficult
```

In cases like this it is generally not possible to bypass the problem using some sort of data transformation and different work-arounds are needed.[7]

4. WORK-AROUNDS

An obvious option to explore when the (pseudo) maximum likelihood estimates exist but convergence is not achieved with the default options is to try one of the different optimization methods offered by the `poisson` command. However, for instance in the third example in Section 3, none of the methods available leads to satisfactory results. Indeed, in that case, with the `NR` and the `BHHH` options the algorithm fails to converge and with the `DFP` and the `BFGS` options the algorithm converges to a result that is far from the optimum. Alternatively, one can ensure convergence just by relaxing the convergence criteria. This, however, is a risky option because the algorithm may be stopped too soon, therefore not delivering the desired (pseudo) maximum likelihood estimates. This is what happens, for instance, in the second example in the previous section when the `nonrtol` option is used.

A simple work-around that often (but by no means always) works is to use the `glm` command with the options `family(poisson) link(log) irls`. Indeed, the iterated re-weighted least squares algorithm provided by the `glm` command appears to be much more stable than the algorithms available in the `poisson` command and it produces the correct results in the three examples presented in Section 3.

To facilitate the estimation of Poisson regressions while STATA does not improve the reliability of `poisson`, we have written the `ppml` command which checks for the existence of the (pseudo) maximum likelihood estimates and offers two methods to drop regressors

[7] Of course, the researcher may want to reconsider the specification being used.

that may cause the non-existence of the estimates. Estimation is then implemented using the `glm` method and `ppml` warns if the variables have large values that are likely to create numerical problems or if there are signs that the convergence is spurious.[8] Further details on `ppml` can be found in the corresponding help file.

5. CONCLUSIONS

In this note we have illustrated some shortcomings of the `poisson` command in STATA. We believe that should be relatively easy to update the command so that it checks for the existence of the Poisson regression estimates and is more resilient to numerical problems.

While an upgraded version of `poisson` is not available, practitioners can use our `ppml` command, which checks for the existence of the estimates before trying to estimate a Poisson regression, and provides several warnings about possible convergence problems.

REFERENCES

Albert, A. and Anderson, J.A. (1984). "On the existence of maximum likelihood estimates in logistic models," *Biometrika*, 71, 1-10.

Cameron, A.C. and Trivedi, P.K. (1998). *Regression analysis of count data*, Cambridge, MA: Cambridge University Press.

Gourieroux, C., Monfort, A. and Trognon, A. (1984). "Pseudo maximum likelihood methods: Applications to Poisson models," *Econometrica*, 52, 701-720.

Haberman, S.J. (1973). "Log-linear models for frequency data: sufficient statistics and likelihood equations," *Annals of Statistics*, 1, 617-632.

Manning, W.G. and Mullahy, J. (2001). "Estimating Log Models: To Transform or Not to Transform?," *Journal of Health Economics* 20, 461-494.

[8] A tell-tale sign that the convergence is spurious is that some zero observations of y are "perfectly predicted"; in the second example in Section 2 the values of $\exp\left(x_i'\hat{\beta}\right)$ for $y = 0$ vary between 3.80E−09 and 4.20E−09.

Santner T.J. and Duffy, E.D. (1986). "A note on A. Albert and J.A. Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models," *Biometrika*, 73, 755-758.

Santos Silva, J.M.C. and Tenreyro, S. (2006), "The log of gravity," *The Review of Economics and Statistics*, 88, 641-658.

Santos Silva, J.M.C. and Tenreyro, S. (2010), "On the existence of the maximum likelihood estimates in poisson regression," *Economics Letters*, 107, 310-312.

Tobin, J. (1958). "Estimation of relationships for limited dependent variables," *Econometrica*, 26, 24–36.

Winkelmann, R. (2008). *Econometric analysis of count data*, 5$^{\text{th}}$ ed., Berlin: Springer-Verlag.